Prise en compte de l'imperfection des tags pour la classification sémantique d'images

Amel Znaidia, Hervé Le Borgne, Céline Hudelot, Adrian Popescu

Prise en compte de l'imperfection des tags pour la classification sémantique d'images

Amel Znaidia[1,2] Hervé Le Borgne[2] Céline Hudelot[1] Adrian Popescu[2]

[1] Ecole Centrale Paris, Laboratoire de Mathématiques Appliquées aux Systèmes
[2] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette CEDEX, France

amel.znaidia@cea.fr , herve.le-borgne@cea.fr , celine.hudelot@ecp.fr , adrian.popescu@cea.fr

Résumé

*L'annotation d'images consiste à décrire le contenu des images en utilisant un nombre fini de **concepts** fixés a priori. En pratique, nous utilisons deux modalités pour cela : l'image et les tags utilisateurs qui les accompagnent. Cependant, ces tags sont en général imparfaits et seulement une partie est pertinente vis-à-vis du contenu de l'image. Dans ce travail, nous nous intéressons à la prise en compte de l'imperfection des tags en vue de leur utilisation pour l'amélioration de la performance des systèmes d'annotation. Nous proposons un système de classification multimodale qui prend en compte l'imperfection des tags. L'amélioration de ≈ 8 % de classification obtenue sur la base d'images VCDT (Visual Concept detection Task) de la campagne d'évaluation ImageClef2011 montre l'intérêt de cette modélisation.*

Mots Clef

Annotation d'images, classification supervisée, imperfection des tags, similarité sémantique.

Abstract

*Image annotation consists in describing the image content according to a finite number of a priori fixed **concepts**. In practice, we use two modalities for this : image and user-tags. However, these tags are generally imperfect and only a part of them are related to the image content. In this work, we are interested in taking into account tag imperfection to improve the performance of annotation systems. Our experimental results on ImageClef2011 PhotoAnnotation benchmark dataset show the interest of this modeling, leading to an improvement of almost ≈ 8 % compared to the visual information only.*

Keywords

Image annotation, supervised classification, tag imperfection, semantic similarity.

1 Introduction

Les grandes collections multimédia, notamment celles disponibles sur le web, nécessitent des outils permettant de les fouiller et de les organiser en fonction de leur contenu. En pratique, celui-ci peut être défini par un ensemble fini de *concepts*, qu'il s'agit d'identifier. Les documents de ces collections sont composés, par définition, de plusieurs « mono-média », qui sont autant de sources d'information. Par exemple, la plupart des sites comme Flickr [1] ou Picasa [2], permettent aux utilisateurs d'annoter leurs images avec des étiquettes (ou tags). Ces tags sont donc censés représenter le contenu des images au sens des utilisateurs. Le contenu est également reflété par d'autres sources d'information, notamment les pixels de l'image. Néanmoins, une source d'information donnée peut être imparfaite. C'est typiquement le cas des tags, ce qui entre en conflit avec la motivation principale des utilisateurs qui est de rendre leurs photos accessibles au grand public [1]. En particulier, Kennedy *et al.* [15] ont montré que les tags produits par les utilisateurs sur Flickr sont en général imparfaits et seulement 50% sont effectivement reliés au contenu de l'image. En effet, les tags ne sont pas imparfaits dans l'absolu mais par rapport à l'usage [9], dans notre cas l'annotation sémantique d'images. La figure 1 montre une telle imperfection

Tags :
BMW E92
BMW 325i
E92 M
BMW 3 series
Coupe
Canon EOS 350D
Xabier Martinez
Auto
Car
coche

Concepts visuels :
Building, city_life, outdoor, sky, clouds, day, no_blur, vehicle, no_person, visual_arts, natural, natural_illumination, car.

FIGURE 1 – Un exemple d'une image de Flickr avec les tags utilisateurs associés et l'ensemble de concepts visuels qui constituent la vérité terrain.

1. http ://www.flickr.com
2. http ://picasa.google.com

sur l'exemple d'une image annotée manuellement provenant de Flickr. Si les tags « auto, car, coche » sont pertinents pour l'image, le tag « coupe » est imprécis du fait de sa synonymie, tandis que d'autres tags comme « outdoor, sky, clouds, building » qui peuvent être utilisés pour décrire le contenu de l'image sont manquants. Ces aspects d'imprécision, d'incertitude et d'incomplétude limitent significativement l'utilisation de cette source d'information textuelle pour la recherche ou l'annotation d'images.

Des travaux récents utilisent les tags pour l'amélioration de la classification d'images en concepts visuels. Guillaumin *et al.* [10] utilisent un noyau linéaire pour modéliser les tags les plus fréquents. Kawanabe *et al.* [14] proposent d'améliorer la méthode de [10] en exploitant l'information de co-occurrence entre les tags via une marche aléatoire. Notre approche utilise aussi les tags pour l'amélioration de la performance de classification mais diffère de ces travaux par la prise en compte de l'imperfection des tags. Nous exploitons la similarité sémantique entre les tags et les concepts visuels pour modéliser cette imperfection. Cette information est ensuite utilisée en entrée d'un classifieur capable de gérer un problème multiclasse. Nous la combinons à une classification classique de la modalité visuelle et montrons que la prise en compte de l'imperfection des tags permet une amélioration notable du taux de classification final.

Le reste de cet article est organisé comme suit. Dans la section 2 nous présentons les travaux connexes dans le domaine de la classification multimodale d'images. Nous proposons notre méthode pour la prise en compte de l'imperfection des tags dans la section 3. La section 4 introduit la base de test et les mesures d'évaluation utilisées pour évaluer notre approche, puis nous discuterons les résultats obtenus dans la section 5. Enfin, nous terminerons par la conclusion et nos perspectives dans la section 6.

2 Etat de l'art

Dans cette section, nous présentons d'une part les travaux connexes en fusion d'information dans le contexte de la classification d'images et d'autre part la modélisation de l'imperfection des tags. Enfin, nous introduisons les mesures de similarité sémantique utilisées dans notre approche.

2.1 Fusion multimodale

Dans la littérature, nous distinguons deux types de stratégies de fusion de données multimodales : la fusion de décisions (ou fusion tardive) et la fusion de descripteurs (ou fusion précoce). La fusion précoce implique d'effectuer le choix d'une méthode pour fusionner les descripteurs et obtenir ainsi un vecteur multimodal. La plus simple, qui est assez largement utilisée, consiste à simplement concaténer les vecteurs uni-modaux [23]. Toutefois, d'autres modèles plus élaborés ont été proposés [21] dans la littérature. A l'opposé, la fusion de décisions consiste à traiter chaque modalité séparément et à fusionner les décisions prises pour chacune de ces modalités [7]. Dans ce cas, nous supposons que les systèmes mono-modaux sont efficaces et que la combinaison des décisions respectives de différentes modalités peut être bénéfique. La technique d'agrégation la plus simple utilisée dans la littérature est la moyenne des scores [7] mais d'autres techniques plus élaborées ont été proposées [3]. L'état de l'art montre que les approches de fusion tardive ont surpassé les approches de fusion précoce [18].

2.2 Modélisation des imperfections des tags

Comme illustré avec l'exemple de la figure 1, les tags associés aux images sur le web sont imparfaits. Cette imperfection a été récemment étudiée dans la littérature en vue d'améliorer la recherche ou l'annotation automatique d'images. Liu *et al.* [17] proposent de reclasser les tags qui accompagnent les images selon un score de pertinence basé sur une estimation de probabilité et raffiné par une marche aléatoire dans un graphe de similarité entre les tags. Sun and Bhowmick [25] proposent un score de clarté pour évaluer l'efficacité d'un tag à décrire le contenu de l'image. Les tags utilisés couramment pour annoter des images visuellement similaires ont un score de clarté plus élevé. Dans [26], les auteurs proposent de déterminer l'ambiguïté des tags et de proposer de nouveaux tags. Cette ambiguïté est basée sur la co-occurence d'un ensemble de tags dans deux contextes différents. Jin *et al.* [13] ont proposé de fusionner les similarités sémantiques par la règle de Dempster-Shafer pour supprimer des mots clés non pertinents. Récemment, Hong *et al.* [11] ont proposé une plateforme collaborative pour l'annotation et la recherche d'images où l'incertitude des tags est déterminée manuellement par les utilisateurs et est modélisée par une valeur de confiance entre 0 et 1. Cependant cette approche manuelle est difficile à appliquer dans des sites de partages de photos comme Flickr. Cela demande aux utilisateurs un effort cognitif supplémentaire non négligeable qui le distrait de sa tâche principale : identifier les tags les plus pertinents. Toutefois, une telle surcharge cognitive devrait être réduite.

2.3 Similarité sémantique

La similarité sémantique entre concepts a été largement étudiée dans la littérature dans le domaine du texte [5]. Dans notre travail, nous avons utilisé deux types de distances pour calculer la similarité entre les tags et les concepts visuels. La première est basée sur la mesure de Wu-Palmer [28] et la deuxième est basée sur les réseaux sociaux [20].

– **La mesure de Wu-Palmer**
  Cette mesure utilise la ressource *WordNet* [8]. Cette dernière peut être considérée comme une hiérarchie sémantique où chaque nœud représente un concept du monde réel. Chaque nœud est composé d'un ensemble de synonymes représentant le même concept, cet ensemble s'appelle *synset*. Les *synsets* sont reliés par des arcs qui décrivent les relations entre les différents concepts. Dans *WordNet*, cette mesure entre deux *synsets* $syns_1$ et

$syns_2$ est définie par :

$$sim_{wup}(syns_1, syns_2) = \frac{2 * depth(lcs(syns_1, syns_2))}{depth(syns_1) + depth(syns_2)} \quad (1)$$

où $lcs(syns_1, syns_2)$ représente le plus petit ancêtre commun des deux *synsets* $syns_1$ et $syns_2$ dans la taxonomie de *WordNet* et $depth(s)$ représente la longueur du chemin reliant s au $Root$ de la taxonomie.

– **La mesure de Popescu [20]**
Cette mesure est basée sur l'information relationnelle entre les tags dans le contexte des réseaux sociaux. Pospescu *et al.* [20], définissent la relation sociale entre deux tags t_i et t_j par :

$$SocRel(t_i, t_j) = users(t_i, t_j) * log(\frac{|users_{collection}|}{users_{collection(t_i)}}) \quad (2)$$

où $users(t_i, t_j)$ représente le nombre d'utilisateurs distincts qui associent le tag t_i au tag t_j parmi les premiers résultats retournés par Flickr pour une requête du tag t_j. $|users_{collection}|$ représente le nombre total d'utilisateurs de la collection, $users_{collection(t_i)}$ le nombre d'utilisateurs qui ont utilisé t_i comme tag pour leur photos. Pour tout tag T_k, nous identifions les tags t_x de Flickr les plus proches en utilisant l'équation (2). Ensuite, nous calculons le modèle de Flickr qui servira par suite à définir la similarité de l'équation (5). Ainsi, le modèle Flickr d'un tag T_k est défini par :

$$M_{Flickr}(T_k) = \cup_{x=1}^{N} (\frac{SocRel(t_x, T_k)}{SocRel(t_1, T_k)}, t_x) \quad (3)$$

où N représente le nombre de tags socialement reliés sur Flickr au sens de (2). Le tag t_1 représente le tag le plus proche de T_k utilisé pour la normalisation.

3 Modèle proposé

Dans cette section, nous présentons d'abord une vue globale de notre approche puis nous détaillons plus précisément comment l'imperfection des tags est prise en compte. Le modèle de classification multimodale d'images en concepts visuels consiste à traiter chaque modalité séparément pour l'étape de la classification et à fusionner ensuite les scores de décisions finaux obtenus pour chaque modalité comme le montre la figure 2. Une description de chaque modalité est d'abord construite pour chaque document multimédia et ces descriptions sont utilisées en entrée de classifieurs multiclasses.

Plus formellement, soient $C = \{C_1, ..., C_K\}$ l'ensemble des concepts visuels de classification, $I = \{I_1, ..., I_{K'}\}$ l'ensemble des images de notre collection et $T = \{T_1, ..., T_{K'}\}$ l'ensemble de tags utilisateurs qui les accompagne. $T_i =< t_{i,1}, ..., t_{i,p} >$ représente l'ensemble de tags associés à une image I_i. Nous disposons d'un ensemble d'apprentissage $A = \{(X_1, y_1), ..., (X_N, y_N)\}$ ou $X_i = (x_i^v, x_i^t)$ représente le descripteur visuel (x_i^v) et le descripteur textuel (x_i^t)

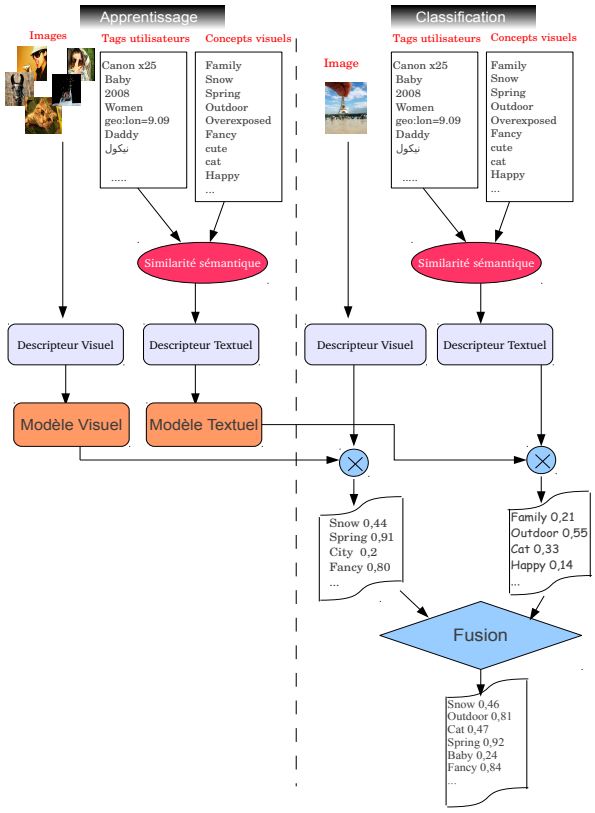

FIGURE 2 – Vu globale du modèle de classification multimodale proposé : 1) chaque modalité est traitée séparément pour l'étape de la classification 2) fusion des scores de décisions finaux obtenus pour chaque modalité.

de l'image I_i et $y_i =< C_1, ..., C_l >$ représente l'ensemble de concepts visuels constituant sa vérité terrain. Pour un document de test (I_j, T_j), notre objectif est de prédire l'ensemble de concepts visuels y_j.

3.1 Description visuelle

Les images sont décrites par cinq descripteurs globaux, rendant compte des couleurs et des textures des images. Concernant la couleur, nous avons utilisé deux histogrammes couleur dans l'espace RGB quantifié sur quatre niveaux (taille $4^3 = 64$), un second sur cinq niveaux (taille $5^3 = 125$) et un troisième dans l'espace couleur HSV quantifié sur 5 bits (taille $5^3 = 125$). Nous avons également ajouté un descripteur couleur prenant en compte la cohérence spatiale des pixels (taille $4^3 = 64$) [24]. Concernant la texture, nous avons calculé un descripteur LEP (local edge pattern, motifs des contours locaux) fournissant un histogramme à 512 composantes [4].

Le descripteur final est la concaténation de ces cinq descripteurs, formant un vecteur de taille 890. Une telle description « globale » est relativement adaptée à la description de scènes, mais semble néanmoins supplantée par les approches à base de descripteurs locaux. Nous avons donc considéré également un tel descripteur. Des descripteurs

SIFT sont extraits selon une grille dense tous les 6 pixels puis codés selon la méthode de [12]. Considérant un dictionnaire de mot visuel de taille $K = 1024$, un mot visuel est codé selon sa « saillance locale » prenant en compte ses 5 plus proches voisins dans le dictionnaire. On ne retient que le code maximal de chaque composante $1 \dots K$ (*maximum pooling*). Le descripteur est calculé pour une pyramide spatiale [16] à deux niveaux, conduisant à un descripteur final de taille $K \times (1 + 4 + 16) = 21504$.

3.2 Description textuelle

Comme Kawanabe *et al.* [14] et Guillaumin *et al.* [10], notre objectif est d'utiliser les tags utilisateurs pour l'amélioration du taux de classification. Comme déjà évoqué en introduction, cette information est imparfaite et nous proposons d'exploiter la similarité sémantique entre les tags et les concepts visuels pour modéliser l'imperfection des tags. L'idée est de projeter les tags dans l'espace conceptuel d'indexation $C = \{C_1, \dots C_K\}$ où les C_i sont les concepts visuels de classification. Nous associons à chaque tag un ou plusieurs concepts visuels selon les similarités entre le tag et les concepts visuels. Le concept associé à plusieurs tags est jugé pertinent pour décrire le contenu de l'image comme illustré sur la figure 3.

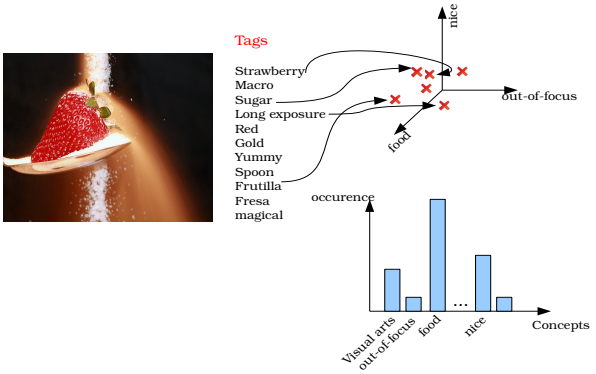

FIGURE 3 – Un exemple d'image avec les tags correspondants. Les tags *« strawberry, sugar, spoon, frutella, fresa »* seront associés au concept *« food »*, jugé donc pertinent pour annoter l'image.

Cet espace de concept peut être vu comme un espace réduit par analyse sémantique latente [6] dont les thèmes (*topics*) sont les concepts visuels de classification.
Après la correspondance des tags aux concepts visuels, nous obtenons un document textuel. Nous représentons donc chaque document suivant le modèle de Salton [22] *i.e. tfidf*, comme un vecteur de poids $(w_{1,j}, ..., w_{i,j}, ..., w_{|C|,j})$ où $w_{i,j}$ représente le poids d'un concept C_i dans une image I_j.
Dans le modèle classique de *tfidf*, nous procédons à une affectation dure (*hard assignment*) pour déterminer la présence ou l'absence d'un concept (1 ou 0). En outre, les utilisateurs de Flickr n'utilisent pas généralement les mêmes concepts visuels de classification pour annoter leurs images. Ainsi, il semble plus approprié de procéder à une affectation souple (*soft assignment*) où un tag sera associé à un concept avec une valeur de confiance. Contrairement aux approches de l'état de l'art où la mesure de confiance est associée à un tag, nous associons une mesure de confiance à un concept visuel.
Soit $s_{k,i}$ la similarité sémantique entre un tag t_k et un concept C_i. Nous avons choisi de calculer $s_{k,i}$ de deux manières distinctes. La première se base sur une ressource externe, dans notre cas *WordNet*. Cela représente une utilisation d'une connaissance de sens commun. La deuxième se base sur une information statistique de co-occurrence de tags dans les réseaux sociaux (c.f. section 2.3).
Pour la première mesure, nous avons utilisé la mesure de Wu-Palmer [28] et la ressource *WordNet*. Etant donné que chaque terme peut appartenir à un ou plusieurs *synsets*, il peut avoir plusieurs sens. Ainsi, nous définissons la similarité entre un tag et un concept comme le maximum de similarité entre leurs *synsets* respectifs. Notons $syns(t_k)$ l'ensemble de *synsets* qui contiennent le tag t_k, nous définissons la similarité entre un tag t_k et un concept visuel C_i comme suit :

$$sim_{Wordnet}(t_k, C_i) = max\{sim_{wup}(s_k, s_i) \; | (s_k, s_i) \in syns(t_k) \times syns(C_i)\} \quad (4)$$

Ensuite, nous avons utilisé le modèle Flickr de Pospescu *et al.* [20]. Dans ce contexte, nous définissons la similarité sémantique entre un tag t_k et un concept visuel C_i par :

$$sim_{Flickr}(t_k, C_i) = \frac{< M_{Flickr}(t_k), M_{Flickr}(C_i) >}{||M_{Flickr}(t_k)|| * ||M_{Flickr}(C_i)||} \quad (5)$$

Notons que nous ne procédons pas à un pré-traitement linguistique pour les tags et les concepts visuels. Ainsi, si le tag ou le concept visuel n'est pas présent dans *WordNet* ou Flickr, la similarité sémantique est égale à 0. Dans ce contexte, nous proposons une nouvelle version de $tfidf$ que nous notons "$\widetilde{tfidf}$". Dans cette méthode, nous ajoutons un score de confiance représentant notre certitude de la présence d'un concept. Les coefficient $\widetilde{tf_{i,j}}$ et $\widetilde{idf_i}$ sont donnés par :

$$\widetilde{tf_{i,j}} = \frac{\sum_{k \in T_j} F_\alpha(s_{k,i})}{\sum_{i \in C} \sum_{k \in T_j} F_\alpha(s_{k,i})} \quad (6)$$

$$\widetilde{idf_i} = log(\frac{|I|}{\sum_{j \in I} \frac{\sum_{k \in T_j} F_\alpha(s_{k,i})}{n_{i,j}}}) \quad (7)$$

où T_j représente l'ensemble des tags associés à l'image I_j, C l'ensemble de concepts visuels de classification et I l'ensemble des images de la collection. $n_{i,j}$ représente le nombre d'occurrences du concept C_i dans l'image I_j. F_α représente une fonction strictement croissante définie par :

$$F_\alpha : [0,1] \longrightarrow [0,1], \quad s_{k,i} \longmapsto \begin{cases} 0 & \text{si} \quad s_{k,i} < \alpha \\ s_{k,i} & \text{si} \quad s_{k,i} \geq \alpha \end{cases} \quad (8)$$

Nous ne prenons en compte que les concepts similaires à

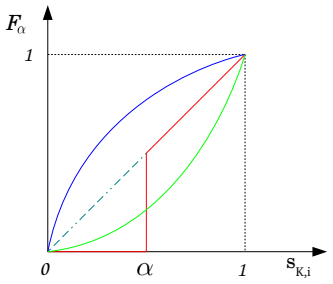

FIGURE 4 – Les fonctions F_α possibles : la fonction concave (en bleu) favorise les similarités tandis que la fonction convexe (en vert) les pénalise.

un certain voisinage. Le seuil de voisinage α est déterminé par validation croisée. Dans le cas où $F_\alpha(s_{k,i})$ est égale à 1, nous retrouvons la formule de $tfidf$ classique. Plusieurs possibilités se présentent pour la fonction F_α comme le montre la figure 4. La fonction concave (en bleu) favorise les similarités tandis que la fonction convexe (en vert) les pénalise. Nous avons choisi la fonction identité à partir d'un certain voisinage (en rouge) pour pénaliser les « petites » similarités que nous supposons produites par des tags imparfaits et donc considérés comme du bruit et nous gardons que les « grandes » similarités.

3.3 Classification et fusion

Pour la classification des deux modalités, nous avons utilisé l'algorithme de partage de caractéristiques (Fast Shared Boosting) [2]. Il permet d'apprendre et de prédire simultanément plusieurs classes. Les scores de prédiction finaux sont obtenus par la moyenne des scores de prédiction de chaque modalité.

4 Bases d'images et évaluation

4.1 Corpus

Nous appliquons notre système multimodal de détection de concepts visuels au corpus de la tâche VCDT (Visual Concept Detection Task) de la campagne internationale ImageClef [19]. Cette tâche correspond à un problème de classification multi-classes multi-labels. Le corpus VCDT contient 8000 images d'apprentissage et 10000 images de test. Ce corpus comprend respectivement 93 et 99 concepts visuels pour VCDT2010 et VCDT2011. Ces concept visuels permettent de décrire une scène « *indoor, outdoor, landscape, mountains ...* », un objet « *dog, car, animal, person, building..* », un événement « *holidays, sport, work ...* », la qualité de l'image « *overexposed, underexposed, blurry...* » ou même des émotions « *funny, calm, nice, melancholic ...* ». Un exemple d'images du corpus avec l'ensemble des concepts visuels qui constituent la vérité terrain est présenté sur la figure 5.

FIGURE 5 – Exemple d'images de la tâche VCDT avec les concepts visuels correspondants.

4.2 Mesures d'évaluations

Pour mesurer les performances de notre système, nous avons utilisé les mesures adoptées par la campagne internationale ImageClef 2011 [19]. Ces mesures sont Mean Average Precision (MAP), Equal Error Rate (l'EER est le taux d'erreur au point de la courbe ROC pour lequel les taux de faux positifs et de faux négatifs sont égaux ; plus il est faible, meilleure est la classification) et Area-under the curve (AUC) pour l'évaluation par concept.

5 Résultats Expérimentaux

Les résultats obtenus par les différentes méthodes sont donnés dans la table 1.

Nous pouvons observer l'amélioration apportée par la prise en compte de l'imperfection des tags. Les méthodes $\widetilde{tfidf}_{wordnet}$ et $\widetilde{tfidf}_{flickr}$ sont plus efficaces que le *tfidf* classique qui obtient un score MAP égal à ≈ 0.14. Les deux classifieurs textuels, basés sur le $\widetilde{tfidf}$ proposé, obtiennent des performances de classification très proches, leur MAP étant respectivement 0.29 et 0.32. La mesure de similarité sémantique basée sur les réseaux sociaux surpasse la mesure basée sur *Wordnet* de ≈ 3 %. Cela peut être expliqué par la nature des tags et la ressource *Wordnet* n'est pas adaptée à l'annotation d'images dans le web [27].

De plus, on observe que le modèle proposé permet d'améliorer le score en cas de fusion de l'information visuelle et de l'information textuelle. Au contraire, avec les méthodes classiques, un conflit entre source d'information visuelle et textuelle (tags) conduit à une diminution des performances en cas de fusion. En fusionnant avec notre méthode, l'élimination des tags donne un gain de 7 à 8 points sur la modalité visuelle seule. La moyenne des scores des

Descripteur	MAP	EER	AUC
$Visuel_{local}$	0.282	0.322	0.736
$Visuel_{global}$	0.300	0.290	0.774
$tfidf_{wordnet}$	0.147	0.473	0.519
$\widetilde{tfidf}_{wordnet}$	0.292	0.356	0.684
$tfidf_{flickr}$	0.136	0.496	0.209
$\widetilde{tfidf}_{flickr}$	0.328	0.305	0.734
$Visuel_{local}$+ $tfidf_{wordnet}$	0.275	0.327	0.727
$Visuel_{local}$+ $\widetilde{tfidf}_{wordnet}$	0.363	0.282	0.782
$Visuel_{local}$+ $tfidf_{flickr}$	0.210	0.367	0.661
$Visuel_{local}$+ $\widetilde{tfidf}_{flickr}$	0.377	0.267	0.799
$Visuel_{global}$+ $tfidf_{wordnet}$	0.291	0.293	0.770
$Visuel_{global}$+ $\widetilde{tfidf}_{wordnet}$	0.372	0.259	0.808
$Visuel_{global}$+ $tfidf_{flickr}$	0.215	0.337	0.693
$Visuel_{global}$+ $\widetilde{tfidf}_{flickr}$	0.383	0.250	0.819
$Visuel_{global}$+ $\widetilde{tfidf}_{flickr}$+ $\widetilde{tfidf}_{wordnet}$	**0.403**	**0.246**	**0.823**

TABLE 1 – Comparaison des différentes méthodes. On distingue l'utilisation de descripteurs visuels globaux et locaux (voir section 3.1)

trois classifieurs (visuel, $\widetilde{tfidf}_{wordnet}$ et $\widetilde{tfidf}_{flickr}$) permet d'obtenir un score MAP de 0.40. Les résultats obtenus par classe montrent que même si l'une des deux modalités représente une meilleure performance, la combinaison donne des meilleures performances pour 91.9 % des concepts sauf pour les concepts « *Desert, Overexposed, Big-group, Aesthetic-impression, Shadow, Birthday, Cat, Old-person* ». Le taux de classification de quelques concepts comme « Desert » reste faible du fait du faible nombre d'images d'apprentissage pour ce concept. Le score MAP de quelques exemples de concepts visuels est donné sur la figure 6. La modalité textuelle améliore la performance de classification pour plusieurs classes comme « *bird, horse, fish, insect, car, bicycle, airplane, winter, snow...* » et échoue pour d'autres comme « Underexposed, overexposed, out-of-focus, abstract ». Ce n'est pas surprenant car généralement les utilisateurs de ces sites de partages utilisent des tags simples et communs pour marquer leurs photos. Plus généralement, ce sont les concepts relatifs à la qualité de l'image et les concepts abstraits liés aux possibles émotions suscitées chez un utilisateur qui sont les plus difficiles à déterminer.

La table 2 présente une comparaison de nos scores MAP avec celle de Kawanabe *et al.* [14] sur le corpus VCDT2010 avec le même protocole d'expérimentation. Ces scores sont obtenus par validation croisée « 20-fold cross-validation » sur les 8000 images d'apprentissage.

La fusion de la modalité visuelle au moyen de notre méthode permet une amélioration de 12 points tandis que la méthode de Kawanabe n'améliore que de 8 points, montrant ainsi l'intérêt de la prise en compte de l'imperfection des tags. On notera que ce gain de 12 points est réalisé

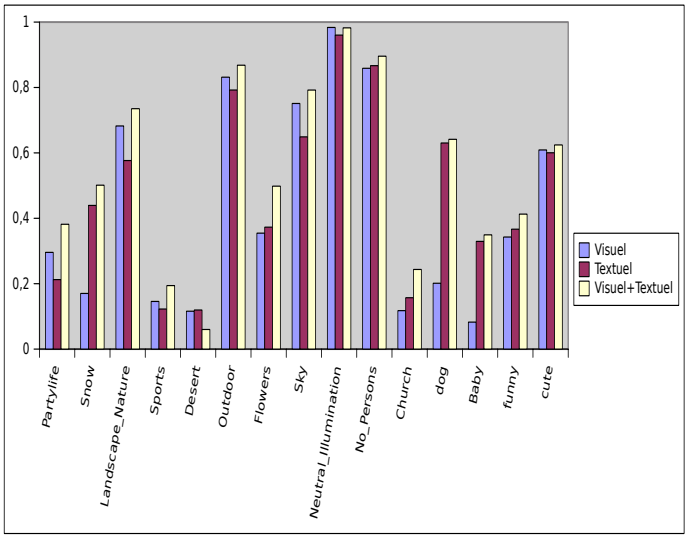

FIGURE 6 – Comparaison du score MAP pour les classifieurs visuel, $\widetilde{tfidf}_{flickr}$ et multimodal.

aussi bien avec les descripteur globaux que locaux. Ce bon résultat est néanmoins à relativiser du fait que notre modalité visuelle est inférieure de 8 à 10 points à celle de [14]. Une explication à cette différence de score « brute » est à chercher dans la nature des descripteurs utilisés : nos descripteurs globaux ont une taille de 890 et nos descripteurs locaux sont de taille 21504. Dans ce dernier cas, nous utilisons de simple SIFT de taille 128 quand Kawanabe les agrège avec quatre descripteurs « SIFT colorés », utilisant ainsi des descripteurs locaux de taille $128 \times 13 = 1664$. Avec autant de codebook de taille 4000 il obtient un descripteur final de taille 160000 soit 7 à 180 fois plus grands que les nôtres.

Méthodes	Taille du descripteur visuel	Visuelle	Gain de fusion
Kawanabe [14]	160000	39.94±1.18	8±2.24
$\widetilde{tfidf} + visuel_{local}$	21504	31.85±2.32	**12±1.34**
$\widetilde{tfidf} + visuel_{global}$	**890**	29.71±2.42	**12±0.92**

TABLE 2 – Comparaison du score MAP par rapport à la méthode de Kawanabe *et al.* [14].

La figure 7 illustre comment les tags aident à améliorer la prédiction des concepts visuels présents dans l'image. Nous comparons les concepts visuels détectés par le classifieur visuel seul et notre méthode multimodale (Visuel+ $\widetilde{tfidf}_{flickr}$) avec celle de la vérité terrain. Les mauvaises détections sont marquées en rouge tandis que les bonnes détections rajoutées par la modalité textuelle sont marquées en vert. Nous observons que la méthode proposée, prenant en compte l'imperfection des tags, permet d'éliminer quelques mauvaises détections et de rajouter des concepts visuels pertinents pour l'annotation de l'image, alors qu'ils sont difficiles à détecter par la modalité visuelle

seule. Par exemple, la première image a pour tag « *dog* », ce qui permet non seulement de détecter ce concept visuel mais aussi d'en inférer d'autres « *Natural, Animals, Cute, Funny* » .

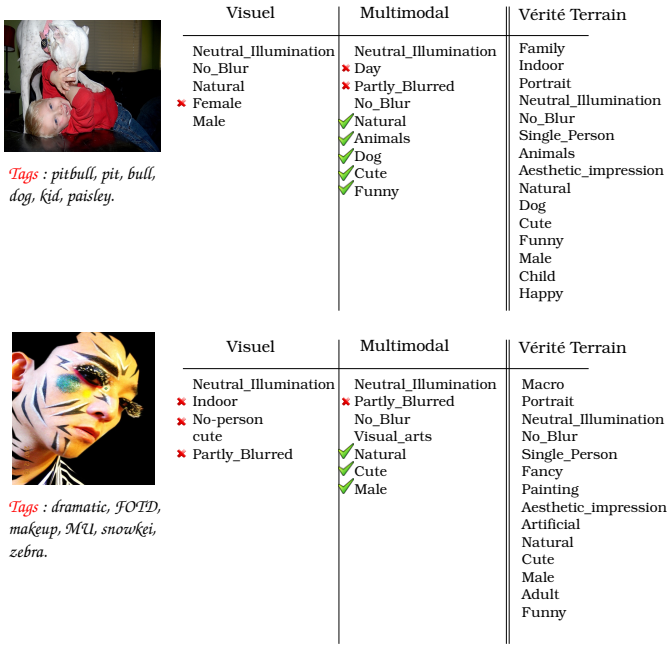

FIGURE 7 – Comparaison des concepts visuels prédits par le classifieur visuel et le système multimodal avec la vérité terrain.

Pour un tag donné, nous n'avons considéré que les concepts visuels sémantiquement similaires à un certain voisinage. C'est une manière de filtrer les concepts visuels dont la similarité sémantique avec un tag est faible. Pour la similarité sémantique basée sur *Wordnet*, cette valeur du voisinage est fixée à la valeur 0.8, ce qui est déterminé comme le seuil permettant d'obtenir le score maximal par validation croisée sur la base d'apprentissage (courbe bleue sur la figure 8). L'expérience est ensuite réalisée entièrement pour différentes valeurs de seuils (courbe rouge de la figure 8), ce qui montre que la valeur déterminée par validation croisée sur la base d'apprentissage permet bien d'obtenir le meilleur score possible sur la base de test.

6 Conclusion

Nous avons proposé un système multimodal de classification d'images en concepts visuels. Cette approche à l'avantage de prendre en compte l'imperfection des tags, en exploitant des similarités sémantiques à l'aide de connaissances externes comme *Wordnet* et les réseaux sociaux.
Les résultats expérimentaux ont démontré l'intérêt de cette approche pour la classification sémantique d'images. De plus l'évaluation a montré une amélioration de $\approx 8\%$ par rapport à la modalité visuelle seule. Les résultats détaillés sur chaque classe montrent que même si l'une des deux

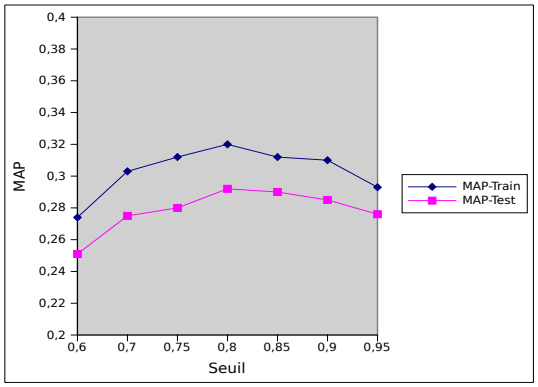

FIGURE 8 – Classification pour différents seuil de voisinage avec la mesure basée sur *Wordnet*. La courbe bleue résulte d'une validation croisée sur la base d'apprentissage, ce qui permet de déterminer le meilleur seuil (0.8). Pour l'expérience complète sur la base de test, le meilleur score possible est bien obtenu pour cette valeur.

modalités représente une meilleure performance, la combinaison donne des meilleurs performances pour 91.9 % des classes.
Néanmoins d'autres types d'imperfections peuvent exister lors de la fusion des scores de prédictions de deux modalités. Nos futurs travaux concerneront la gestion des imperfections pour la modalité visuelle ainsi que sa fusion avec les travaux présents, dans un cadre formel permettant de prendre en compte l'imperfection générale des données.

Remerciements

Projet soutenu par l'attribution d'une allocation doctorale DIGITEO et la Région Ile-de-France.

Références

[1] M. Ames and M. Naaman. Why we tag : motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM.

[2] H. L. Borgne and N. Honnorat. Fast shared boosting : Application to large-scale visual concept detection. In G. Quénot, editor, *International Workshop on Content Based Multimedia Indexing, CBMI*, pages 13–18, Grenoble, France, 2010.

[3] J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. González. Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 359–366, New York, NY, USA, 2010. ACM.

[4] Y.-C. Cheng and S.-Y. Chen. Image classification using color, texture and regions. *Image Vision Computing*, 2003.

[5] C. D'Amato, S. Staab, and N. Fanizzi. On the influence of description logics ontologies on conceptual similarity. In *Proceedings of the 16th international conference on Knowledge Engineering : Practice and Patterns*, EKAW '08, pages 48–63. Springer-Verlag, 2008.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. volume 41, pages 391–407, 1990.

[7] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *MIR '08 : Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179, New York, NY, USA, 2008. ACM.

[8] C. Fellbaum, editor. *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[9] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32 :198–208, April 2006.

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 902 – 909, jun 2010.

[11] Y. Hong and S. Reiff-Marganiec. Towards a collaborative framework for image annotation and search. In *CAiSE Workshops*, pages 564–574, 2011.

[12] Y. Huang, K. Huang, Y. Yu, and T. Tan. Salient coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1753 –1760, june 2011.

[13] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *ACM Multimedia*, pages 706–715, 2005.

[14] M. Kawanabe, A. Binder, C. Muller, and W. Wojcikiewicz. Multi-modal visual concept classification of images via markov random walk over tags. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV)*, WACV '11, pages 396–401, Washington, DC, USA, 2011. IEEE Computer Society.

[15] L. S. Kennedy, S. fu Chang, and I. V. Kozintsev. To search or to label ? : predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258, 2006.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169 – 2178, 2006.

[17] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 351–360, New York, USA, 2009. ACM.

[18] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *ImageCLEF : Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer, Berlin, 2010.

[19] S. Nowak, K. Nagel, and J. Liebetrau. The clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF 2011 working notes*, 2011.

[20] A. Popescu and G. Grefenstette. Social media driven image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 33 :1–33 :8, New York, NY, USA, 2011. ACM.

[21] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, MM '10, pages 251–260, New York, NY, USA, 2010. ACM.

[22] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11) :613–620, November 1975.

[23] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *In ACM Multimedia*, pages 399–402, 2005.

[24] R. O. Stehling, M. A. Nascimento, and A. X. Falcao. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 102–109, McLean, Virginia, USA, 2002.

[25] A. Sun and S. S. Bhowmick. Image tag clarity : in search of visual-representative tags for social images. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 19–26, New York, NY, USA, 2009. ACM.

[26] K. Q. Weinberger, M. Slaney, and R. Van Zwol. Resolving tag ambiguity. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 111–120. ACM, 2008.

[27] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 31–40, New York, NY, USA, 2008. ACM.

[28] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico State University, Las Cruces, New Mexico, 1994.